Journal of Mathematics, Computations, and Statistics (hal. 20 – 31)
Vol. 3. No. 1, April 2020
http://www.ojs.unm.ac.id/jmathcos

Model Regresi Semiparametrik Spline untuk D ata Longitudinal pada Kasus Demam Berdarah Dengue di Kota Makassar

Syafruddin Side[1,a)], Wahidah Sanusi[1,b)], dan Mustati'atul Waidah Maksum[1, c)]

[1]*Jurusan Matematika FMIPA Universitas Negeri Makassar, 90224*

[a)]syafruddin@unm.ac.id
[b)]wahidah.sanusi@unm.ac.id
[c)]mustatiatulwaidahmaksum@gmail.com

Abstrak. *Regresi semiparametrik merupakan model regresi yang memuat komponen parametrik dan komponen nonparametrik dalam suatu model. Pada penelitian ini digunakan model regresi semiparametrik spline untuk data longitudinal dengan studi kasus penderita Demam Berdarah Dengue (DBD) di Rumah Sakit Universitas Hasanuddin Makassar periode bulan Januari sampai bulan Maret 2018. Estimasi model regresi terbaik didapat dari pemilihan titik knot optimal dengan melihat nilai Generalized Cross Validation (GCV) dan Mean Square Error (MSE) yang minimum. Komponen parametrik pada penelitian ini adalah hemoglobin (g/dL) dan umur (tahun), suhu tubuh (℃), trombosit ($\times 10^3 \mu L$) sebagai komponen nonparametrik dengan nilai GCV minimum sebesar 221,67745153 dicapai pada titik knot yaitu 14,552; 14,987; dan 15,096; nilai MSE sebesar 199,1032; dan nilai koefisien determinasi sebesar 75,3% yang diperoleh dari model regresi semiparametrik spline linear dengan tiga titik knot..*
Kata Kunci: *regresi semiparametrik, spline, knot, Generalized C*ross Validation, Demam Berdarah *Dengue.*

Abstract. *Semiparametric regression is a regression model that includes parametric and nonparametric components in it. The regression model in this research is spline semiparametric regression with case studies of patients with Dengue Hemorrahagic Fever (DHF) at University of Hasanuddin Makassar Hospital during the period of January to March 2018. The best regression model estimation is obtained from the selection of optimal knot which has minimum Generalized Cross Validation (GCV) and Mean Square Error (MSE). Parametric component in this research is hemoglobin (g/dL) and age (years), body temperature (℃), platelets ($\times 10^3 \mu L$) as a nonparametric components. The minimum value of GCV is 221,67745153 achieved at the point 14,552; 14,987; and 15,096 knot; MSE value of 199,1032; and the value of coefficient determination is 75,3% obtained from semiparametric regression model linear spline with third point of knots.*
Keywords: *semiparametric regression, spline, knot, Generalized Cross Validation, Dengue Hemorrahagic Fever.*

PENDAHULUAN

Data longitudinal merupakan salah satu bentuk data berkorelasi. Dalam studi longitudinal dimungkinkan untuk mempelajari perubahan respons antar waktu beserta faktor yang mempengaruhi perubahan tersebut, baik pada level populasi maupun level individu. Penentuan pilihan dimensi waktu sangat tergantung pada pertanyaan penelitian yang ingin dijawab atau tujuan penelitian yang ingin dicapai (Poerwanto, & Budiantara, 2014).

Analisis tentang pemodelan data longitudinal dapat dilakukan dengan regresi semiparametrik. Regresi semiparametrik merupakan gabungan antara regresi parametrik dan regresi nonparametrik. Pada pendekatan regresi parametrik diasumsikan bahwa bentuk kurva regresi diketahui berdasarkan informasi sebelumnya (teori). Akibatnya estimator kurva regresi diperoleh dengan mengestimasi parameternya. Pendekatan regresi nonparametrik tidak memberikan asumsi bentuk kurva tertentu ataupun tidak ada informasi mengenai bentuk kurva regresi. Kurva regresi nonparametrik dapat diasumsikan mulus atau *smooth*, sehingga regresi nonparametrik memiliki fleksibilitas yang tinggi (Utami, 2014).

Pendekatan regresi nonparametrik telah banyak dikembangkan, antara lain menggunakan *spline*, *kernel*, polinomial lokal, *wavelet*, dan *fourier*. Salah satu model regresi dengan pendekatan nonparametrik yang sangat sering digunakan untuk melakukan estimasi terhadap kurva regresi adalah regresi *spline* (Adawiyah, 2018). *Spline* adalah salah satu jenis *piecewise polynomial*. Maksud *piecewise polynomial* adalah polinomial yang memiliki sifat tersegmen atau sifat terpotong-potong. Model polinomial dengan sifat terpotong-potong menyebabkan *spline* memiliki fleksibilitas yang lebih tinggi dari model polinomial biasa, sehingga menyebabkan regresi *spline* dapat menyesuaikan diri secara lebih efektif terhadap karakteristik lokal suatu fungsi data (Yani, 2017).

Beberapa penelitian telah menggunakan model regresi semiparametrik (Abdy, 2009; Yani, 2017; Laome, 2009). Abdy (2009) memodelkan regresi semiparametrik dengan menggunakan pendekatan *Generalized Estimating Equation* (GEE), Yani (2017) memodelkan regresi semiparametrik dengan *spline truncated* untuk data pasien DBD, Laome (2009) memodelkan regresi semiparametrik *spline* untuk data longitudinal pada kadar CD4 penderita HIV. Pada penelitian ini, akan memodelkan regresi semiparametrik untuk data longitudinal dengan pendekatan *Generalized Estimating Equation* (GEE). Hasil pemodelan regresi semiparametrik pada data longitudinal diperoleh dari data kasus Demam Berdarah Dengue (DBD).

KAJIAN PUSTAKA

Demam Berdarah Dengue (DBD)

Demam Berdarah Dengue (DBD) adalah penyakit febril akut yang ditemukan di daerah tropis, dengan penyebaran geografis yang mirip dengan malaria. Penyakit ini disebabkan oleh salah satu dari empat serotipe virus dari genus *Flavivirus*, famili *Flaviviridae*. Terdapat tiga faktor pemegang peran dalam penularan infeksi virus dengue yaitu manusia, virus, dan vektor perantara (Yani, 2017). Demam berdarah umumnya lamanya sekitar enam atau tujuh hari dengan puncak demam yang lebih kecil terjadi pada akhir masa demam (Utami, 2014).

Analisis Regresi Parametrik

Analisis regresi merupakan suatu studi yang digunakan untuk melihat ketergantungan atau hubungan antara suatu variabel respons (y_i) pada satu atau lebih variabel prediktor (x_i). Hubungan antara y_i dengan x_i dapat dinyatakan dalam persamaan berikut (Yani, 2017) :

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n \quad (1)$$

dalam hal ini y_i adalah nilai variabel respons dalam pengamatan ke-i, x_i adalah variabel prediktor dalam amatan ke-i, $f(x_i)$ adalah regresi yang telah diketahui bentuknya, ε_i adalah error (kesalahan) acak yang diasumsikan identik, independen dan berdistribusi normal dengan mean nol dan variansi σ^2 atau $N(0, \sigma^2)$ d an n adalah banyaknya amatan.

Model antara dua atau lebih variabel prediktor ($x_1, x_2, \dots, x_k$) dengan variabel respons (y_i) secara umum dapat ditulis sebagai berikut (Gusti, 2011) :

$$\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{x_{i1}} + \beta_2 \mathrm{x_{i2}} + \cdots + \beta_k \mathrm{x_{ik}} + \varepsilon_i, \mathrm{i} = 1,2, \dots, \mathrm{n} \quad (2)$$

dimana y_i adalah variabel respon, $x_{i1}, x_{i2}, \dots, x_{ik}$ adalah variabel predictor, $\beta_0, \beta_1, \dots, \beta_k$ adalah parameter yang tidak diketahui, dan ε_i adalah error acak yang diasumsikan identik, independen dan berdistribusi normal dengan mean nol dan varians σ^2.

Regresi Nonparametrik Spline

Secara umum model regresi nonparametrik dapat dituliskan sebagai berikut (Gusti, 2011) :

$$y_i = f(t_i) + \varepsilon_i, i = 1,2,3, \dots, n \quad (3)$$

dimana y_i adalah variabel respon, $f(t_i)$ adalah regresi yang tidak diketahui bentuk atau polanya. ε_i adalah error acak yang diasumsikan identik, independen, dan berdistribusi normal.

Regresi *spline* merupakan salah satu teknik estimasi dalam regresi nonparametrik dengan model polinomial yang memiliki sifat tersegmen yang mulus. Apabila regresi $f(t_i)$ pada persamaan (3) dihampiri fungsi *spline*, maka untuk mengestimasi $f(t_i)$ dapat digunakan model regresi *spline*. Secara umum model regresi *spline* pada suatu fungsi dengan orde m dapat dinyatakan sebagai berikut (Adawiyah, 2018) :

$$f(t_i) = \sum_{j=0}^{m} \beta_j t_i^j + \sum_{l=1}^{p} \beta_{(m+l)} (t_i - k_l)^m \quad (4)$$

dengan $(t_i - k_l)^m = \begin{cases} (t_i - k_l)^m; & t_i \geq k_l \\ 0 & ; \ t_i < k_l \end{cases}$

dimana $f(t_i)$ adalah fungsi regresi spline, $k_1, k_2, \dots, k_k$ adalah titik knot, t adalah variabel prediktor, dan β adalah konstanta.

Regresi Semiparametrik

Model regresi semiparametrik merupakan gabungan dari model regresi parametrik dan regresi nonparametrik (Gusti, 2011). Adapun model regresi semiparametrik adalah

$$Y_i = X_i \beta + f(t_i) + \varepsilon_i, i = 1,2,3, \dots, n \quad (5)$$

dimana Y_i adalah nilai variabel respon dalam amatan ke-i, X_i adalah variabel prediktor yang berhubungan secara parametrik dengan variabel respon Y_i. β ($\beta_0, \beta_1, \dots, \beta_h$) merupakan parameter koefisien regresi. Sementara itu $f(t_i)$ adalah fungsi regresi yang tidak diketahui bentuk polanya, dan ε_i adalah error acak yang diasumsikan identik, independen, dan berdistribusi normal dengan mean nol dan varians σ^2.

Pendugaan Parameter *Generalized Estimating Equation* (GEE)

Generalized Estimating Equation (GEE) merupakan perkembangan dari *Generalized Linear Model* (GLM) yang diperkenalkan oleh Liang dan Zeger (1986) yang digunakan untuk menduga parameter model berdasarkan data yang mengandung autokorelasi dan data yang tidak menyebar normal (Handayanti, 2015). Penduga parameter β dapat diperoleh dengan menyelesaikan persamaan berikut (Danardono, 2015) :

$$\sum_{i=1}^{n} D_i^T V_i^{-1} (Y_i - \mu_i) = 0 \quad (6)$$

dengan $D_i = \frac{\partial \mu_i}{\partial \beta}$. Matriks D_i yaitu matriks yang berisi turunan μ_i terhadap komponen β. Sedangkan V_j merupakan matriks ragam-peragam Y_{ij} yang berukuran $n_j \times n_j$ pada subyek ke-i, yakni :$V_j = \left[\emptyset A_j^{1/2} R(\alpha) A_j^{1/2}\right]$. Pada penelitian ini, pendugaan parameter β pada persamaan (5) akan diperoleh menggunakan pendekatan parameter GEE dengan menyelesaikan persamaan (6).

Pemilihan Titik Knot Optimal

Titik knot merupakan titik perpaduan bersama yang memperlihatkan terjadinya perubahan perilaku dari fungsi *spline* pada interval-interval yang berbeda sehingga kurva yang terbentuk tersegmen pada titik tersebut. Pada penetuan model regresi *spline* dapat dilakukan dengan melihat nilai *Generalized Cross Validation* (GCV) yang minimum. Adapun rumus untuk menghitung GCV adalah sebagai berikut (Yani, 2017) :

$$GCV(K) = \frac{MSE(K)}{(n^{-1}tr[I-A(K)])^2} \tag{7}$$

dengan MSE(K) = $n^{-1}\sum_{i=1}^{n}[Y_i - \hat{f}(t_i)]^2$, K adalah titik knot $(K_1, K_2, K_3, \dots, K_n)$, $\hat{f}(t_i) = t(t^1t)^{-1}t'Y$, n adalah jumlah data, I adalah matriks identitas, $A(K) = t(t^1t)^{-1}t'$ dan $\hat{Y}_i = A(K)Y$.

Titik knot yang akan digunakan pada penelitian ini adalah K=1,2,3. Kemudian akan dipilih satu titik knot yang memiliki nilai GCV paling minimum yang merupakan titik knot terbaik untuk model regresi *spline.*

Pengujian Parameter Model

Pengujian parameter model secara serentak (simultan) dilakukan dengan menggunakan uji F. Adapun hipotesis pada uji F ini adalah sebagai berikut (Gusti, 2011) :

$H_0: \beta_j = 0, \ \forall j = 0,1,2, \dots, n$

H_1: minimal terdapat satu $\beta_j \neq 0$, untuk suatu $j = 0,1,2, \dots, n$

Adapun statistik uji yang digunakan pada uji F adalah sebagai berikut :

$$F = \frac{KT_{Reg}}{S^2}$$

dengan keputusan jika $F_{hitung} > F_{tabel}$ maka tolak H_0dan terima H_1.

Sedangkan pengujian parameter model secara parsial (individu) yaitu dengan menggunakan pendekatan GEE dapat dilihat berdasarkan nilai signifikansi atau *p-value* dari uji Wald. Apabila $p - value$ kurang dari taraf nyata (α) maka parameter tersebut signifikan terhadap variabel respon.

Pengujian Asumsi Residual

Residual (*goodness of fit)* dari suatu model regresi harus memenuhi asumsi *IIDN* $(0,\sigma^2)$ yaitu identik, independen, dan berdistribusi normal. Pemeriksaan asumsi homogenitas atau uji asumsi identik dapat dilakukan dengan menggunakan uji Glejser. Uji Glejser mempertimbangkan regresi nilai $|e_i|$ terhadap variabel X yang dianggap berhubungan dekat dengan varians heteroskedastisitas α_i^2. Sementara itu asumsi residual independen dapat dilakukan dengan menggunakan uji d Durbin-Watson, adapun uji d Durbin-Watson dapat dirumuskan sebagai berikut :

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{8}$$

Sedangkan pengujian asumsi normalitas dapat dilakukan dengan melakukan uji Anderson-Darling. Adapun uji Anderson-Darling dapat dirumuskan sebagai berikut :

$$A^2 = -n - S \tag{9}$$

Dengan $S = \frac{1}{n}\sum_{i=1}^{n}[2i-1][\log(F(Z_i)) + \log(1 - F(Z_{n+1-i}))]$ adalah simpangan baku data, X_i adalah data ke-i yang telah diurutkan, $\bar{X}$ adalah rata-rata data, $F(Z_i)$ adalah nilai fungsi distribusi

kumulatif normal baku di Z_i, n adalah ukuran sampel, dan $F_0(X)$ adalah fungsi distribusi kumulatif teoritis (Yani, 2017).

TABEL 1. Tabel aturan keputusan uji d Durbin-Watson

Hipotesis nol	Keputusan	Jika
Tidak ada autokorelasi positif	Tolak	$0 < d < d_L$
Tidak ada autokorelasi positif	Tidak ada keputusan	$d_L < d < d_u$
Tidak ada autokorelasi negatif	Tolak	$4 - d_L < d < 4$
Tidak ada autokorelasi negatif	Tidak ada keputusan	$4 - d_U \leq d < 4 - d_L$
Tidak ada autokorelasi positif atau negatif	Terima	$d_U \leq d < 4 - d_U$

Data Longitudinal

Data longitudinal adalah data pengamatan berulang pada unit eksperimen, berbeda dengan data *cross section* yaitu data dari masing-masing individu diamati dalam sekali waktu. Ada beberapa keuntungan dari studi mengenai data longitudinal dibandingkan dengan data *cross section*. Pertama, studi longitudinal lebih *powerful* dari studi *cross section* untuk sejumlah subjek yang tetap. Dengan kata lain, untuk memperoleh kekuatan uji statistik yang sama, studi longitudinal membutuhkan subjek yang lebih sedikit. Kedua, dengan jumlah subjek yang sama, hasil pengukuran *error* menghasilkan penaksir efek perlakuan yang lebih efisien dari data *cross section*. Ketiga, data longitudinal mampu menyediakan informasi tentang perubahan individu, sedangkan data *cross section* tidak (Laome, 2009).

Data pengamatan berulang yang digunakan dalam penelitian ini dilakukan terhadap pasien Demam Berdarah Dengue. Perubahan pasien DBD yang diamati adalah variabel-variabel yang mempengaruhi lama kesembuhan pasien seperti kadar trombosit, suhu tubuh, hemoglobin.

METODE PENELITIAN

Penelitian ini merupakan penelitian terapan. Tujuan penelitian ini untuk mengetahui model regresi semiparametrik *spline* pada data longitudinal. Jenis data yang digunakan adalah data rekam medis pasien DBD yang menjalani rawat inap di Rumah Sakit Unhas Makassar sebanyak 58 sampel periode bulan Januari sampai bulan Maret 2018. Variabel penelitian yang digunakan terdiri dari satu variabel respons yaitu lama kesembuhan dan empat variabel yang diduga berpengaruh yaitu usia, kadar trombosit, hemoglobin, suhu tubuh.
Langkah-langkah analisis data yang digunakan dalam peneltian ini adalah menetapkan komponen parametrik dan komponen nonparametrik, memilih titik *knot* optimal pada komponen nonparametrik, dan memodelkan data. Dalam menganalisis data dilakukan pula pengujian parameter model baik secara serentak maupun individu dan pengujian asumsi residual IIDN.

HASIL PENELITIAN

Estimasi Model Regresi Semiparametrik *Spline* dengan Pendekatan *Generalized Estimating Equation* (GEE)

Diberikan model regresi semiparametrik (10).

$$Y_{ij} = X_{ij}\beta + f(t_{ij}) + \varepsilon_{ij}, i = 1, \dots, n; j = 1, \dots, m \quad (10)$$

Dimana Y_{ij} adalah variabel respon kelompok ke-i untuk pengamatan ke-j, X_{ij} merupakan komponen parametrik dan $f(t_{ij})$ adalah komponen nonparametrik yang merupakan fungsi *spline*

yang tidak diketahui dan $\varepsilon_{ij} \sim NII(0, \sigma^2)$. Berdasarkan fungsi penghubung GEE, bentuk (10) dapat dinyatakan dalam bentuk (11).

$$l(\mu_{ij}) = X_{ij}\beta + f(t_{ij}) \tag{11}$$

Dimana $l(.)$ adalah suatu fungsi penghubung dan $\mu_{ij} = E[Y_{ij}|X_{ij}]$. Parameter β akan diestimasi dengan menggunakan GEE dan fungsi mulus $f(t)$ akan diestimasi dengan memaksimumkan penalized quasi-likelihood. Pendekatan GEE untuk mengestimasi β pada model regresi parametrik akan diuraikan sebagai berikut :

Misalkan $X_i = (X_{i1}, \dots, X_{in})^T$ adalah matriks $n \times p$ yang merupakan nilai dari variabel prediktor untuk subjek ke-i $(i = 1, \dots, n)$. Diberikan persamaan estimasi GEE pada persamaan (12) :

$$\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) = 0 \tag{12}$$

Dimana $\mu_i = E(Y_i)$ yang mempunyai komponen ke-j $\mu_{ij} = X_{ij}\hat{\beta}$, $D_i = \frac{\partial \mu_i}{\partial \beta}$, dan $V_i = \emptyset A_i^{\frac{1}{2}} R_i(\alpha) A_i^{1/2}$, dimana $R_i(\alpha)$ adalah matriks korelasi berukuran $n \times n$, untuk n data berulang untuk satu individu i, dan A_i adalah matriks diagonal berukuran $n \times n$ dengan $V(\mu_{ij})$ adalah elemen diagonalnya, $\emptyset$ adalah suatu parameter dispersi (penyebaran). Berdasarkan bentuk (12) diperoleh parameter $\hat{\beta}$:

$$\hat{\beta} = \left[\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{n} X_i^T V_i^{-1} Y_i\right] \tag{13}$$

Kemudian menggunakan estimasi parameter ke t untuk memperbaharui $\hat{\beta}$ dalam persamaan (14) :

$$\hat{\beta}^{t+1} = \hat{\beta}^{t} - \left[\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{n} X_i^T V_i^{-1} Y_i\right] \tag{14}$$

Selanjutnya, fungsi mulus $f(t)$ akan diestimasi dengan memaksimumkan penalized quasi-likelihood. Definisi fungsi penalized quasi-likelihood adalah (Ibrahim, & Suliadi, 2009) :

$$\Pi = \Phi - \frac{1}{2}\lambda \int_a^b [f''(t)]^2 dt \tag{15}$$

Dimana $\Phi = \sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i)$, $\lambda > 0$ adalah faktor penalized dan $\int_a^b [f''(t)]^2 dt = f^T G f$ dimana G adalah matriks simetris (Ibrahim, & Suliadi, 2009).

Misalkan $Y_i = (Y_{i1}, \dots, Y_{in})^T$ adalah vektor $n \times 1$ yang menyatakan variabel respons, $f = \left[f(t_{(1)}), f(t_{(2)}), \dots f(t_{(q)})\right]^T$. $f(.)$ akan diestimasi dengan memaksimumkan bentuk (15) :

$$\begin{aligned}
\frac{\partial \Pi}{\partial f} &= \sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) - \frac{\partial}{\partial f}\left[\frac{1}{2}\lambda \int_a^b [f''(t)]^2 dt\right] \\
&= \sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) - \lambda G \hat{f} \\
&\Leftrightarrow \sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) - \lambda G \hat{f} = 0 \\
&\Leftrightarrow \hat{f} = \left[\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) - \lambda G\right]^{-1}
\end{aligned} \tag{16}$$

Kemudian menggunakan estimasi parameter ke t untuk memperbaharui $\hat{f}$ dalam persamaan (17):

$$\hat{f}^{t+1} = \hat{f}^{t} + \left[\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) - \lambda G\right]^{-1} \left[\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) - \lambda G \hat{f}^t\right] \tag{17}$$

Deskripsi Data

Gambaran umum data penelitian yaitu lama kesembuhan pasien DBD (hari) sebagai peubah respon (Y) dan peubah bebas (Xi) yaitu umur (tahun), suhu tubuh (℃), trombosit ($\times 10^3 \mu L$), dan hemoglobin (g/dL) diringkas dalam statistika deskriptif pada Tabel 2.

TABEL 2. Statistika deskriptif data pasien Demam Berdarah Dengue (DBD)

Variabel	Ringkasan Statistik			
	Min	**Max**	**Mean**	**StDev**
Lama Kesembuhan (Y)	2,0	9,0	4,776	1,511
Suhu (S)	36,240	38,733	37,050	0,566
Umur (U)	1,0	66,0	22,22	15,60
Trombosit (PLT)	32,3	421,8	147,5	88,2
Hemoglobin (HB)	10,633	15,967	13,636	1,463

Penetuan Komponen Parmetrik dan Komponen Nonparametrik

Tahap awal sebelum memodelkan kasus Demam Berdarah Dengue adalah menentukan variabel parametrik dan nonparametrik dengan melakukan pengujian. Untuk mengetahui data tersebut dalam kelompok parametrik atau nonparametrik, terlebih dahulu akan dilakukan uji normalitas terhadap data dengan cara melihat plot dari data tersebut menggunakan hipotesis berikut :

H_0: $\mu = 0$ (normal)

H_1: $\mu \neq 0$ (tidak normal)

Kriteria untuk menolak atau menerima H_0 berdasarkan *p-value* dinyatakan sebagai berikut :

Jika p-value$< \alpha = 0.05$, maka H_0 ditolak yang berarti variabel tersebut merupakan komponen nonparametrik.

Jika p-value$> \alpha = 0.05$, maka H_0 diterima yang berarti variabel tersebut merupakan komponen parametrik.

Diperoleh hasil uji normalitas pada masing-masing variabel prediktor yaitu p-value < 0,05 pada variabel suhu tubuh, umur, dan trombosit. P-value < 0,05 maka H_0 ditolak yang berarti ketiga variabel tersebut merupakan komponen nonparametrik. Sedangkan p-value > 0,05 pada variabel hemoglobin maka H_0 diterima yang berarti variabel tersebut merupakan komponen parametrik.

Sehingga data pasien Demam Berdarah Dengue dapat didekati dengan regresi semiparametrik *spline*. Terdapat 1 variabel prediktor yang merupakan komponen parametrik yaitu hemoglobin dan 3 variabel prediktor yang merupakan komponen nonparametrik yaitu Suhu, Umur, dan Trombosit.

Pemodelan Regresi Semiparametrik *Spline*

Berdasarkan penentuan komponen parametrik dan komponen nonparametrik, model regresi semiparametrik *spline* dapat dituliskan dalam persamaan (18).

$$Y_i=\beta_0+\beta_1 X_{1n}+\beta_2 Z_{1n}+\sum_{j=1}^{m}\beta_j(Z_{1n}-k_j)+\beta_3 Z_{2n}+\sum_{j=1}^{m}\beta_j(Z_{1n}-k_j)+\beta_4 Z_{3n}+\sum_{j=1}^{K}\beta_j(Z_{1n}-k_j)+\varepsilon_i \quad (18)$$

Dengan

$$X=\begin{bmatrix}1 & x_{11} & z_{11} & z_{21} & z_{31}\\ 1 & x_{12} & z_{12} & z_{22} & z_{32}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{1(58)} & z_{1(58)} & z_{2(58)} & z_{3(58)}\end{bmatrix}, \beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{bmatrix}$$

$$Z=\begin{bmatrix}(z_{11}-k_1) & (z_{21}-k_2) & (z_{31}-k_3)\\ (z_{12}-k_1) & (z_{22}-k_2) & (z_{32}-k_3)\\ \vdots & \vdots & \vdots\\ (z_{1(58)}-k_1) & (z_{2(58)}-k_2) & (z_{3(58)}-k_3)\end{bmatrix}$$

Dimana X merupakan komponen parametrik, Z adalah komponen nonparametrik, k adalah titik knot.

Penentuan Titik Knot Optimal Regresi Semiparametrik *Spline*

Titik knot yang dicobakan pada penelitian ini sampai tiga titik knot (k =1,2,3), bertujuan agar mempermudah dalam melakukan interpretasi. Pemilihan titik knot optimal yang dicobakan adalah:

1. Titik knot optimal regresi semiparametrik *spline* dengan satu titik knot, diperoleh model (19).

   $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{1n} + \hat{\beta}_2 z_{1n} + \hat{\beta}_3 (z_{1n} - k_1) + \hat{\beta}_4 (z_{2n} - k_2) + \hat{\beta}_5 (z_{3n} - k_3) \quad (19)$$

   Nilai GCV minimum sebesar 244.8255613 dicapai pada titik knot untuk variabel umur (Z_1) 38.68244898, variabel suhu (Z_2) 413.80102041, dan variabel trombosit (Z_3) 15.85782313.

2. Titik knot optimal regresi semiparametrik *spline* dengan dua titik knot, diperoleh model (20).

   $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{1n} + \hat{\beta}_2 z_{1n} + \hat{\beta}_3 (z_{1n} - k_1) + \hat{\beta}_4 (z_{1n} - k_1) + \hat{\beta}_5 z_{2n} + \hat{\beta}_6 (z_{2n} - k_2) + \hat{\beta}_7 (z_{2n} - k_2) + \hat{\beta}_8 z_{3n} + \hat{\beta}_9 (z_{3n} - k_3) + \hat{\beta}_{12} (z_{3n} - k_3) \quad (20)$$

   Nilai GCV minimum sebesar 244.82556133 dicapai pada titik knot untuk variabel umur (Z_1) 38.68244898 dan 38.73333333, variabel suhu (Z_2) 413.80102041 dan 421.75 dan variabel trombosit (Z_3) 15.85782313 dan 15.96666667.

3. Titik knot optimal regresi semiparametrik *spline* dengan tiga titik knot, diperoleh model (21).

   $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{1n} + \hat{\beta}_2 z_{1n} + \hat{\beta}_3 (z_{1n} - k_1) + \hat{\beta}_4 (z_{1n} - k_1) + \hat{\beta}_5 (z_{1n} - k_1) + \hat{\beta}_6 z_{2n} + \hat{\beta}_7 (z_{2n} - k_2) + \hat{\beta}_8 (z_{2n} - k_2) + \hat{\beta}_9 (z_{2n} - k_2) + \hat{\beta}_{10} z_{3n} + \hat{\beta}_{12} (z_{3n} - k_3) + \hat{\beta}_{13} (z_{3n} - k_3) \quad (21)$$

   Nilai GCV minimum sebesar 221.67745153 dicapai pada titik knot untuk variabel umur (Z_1) 38.07183673,38.27537415 dan 38.32625850, variabel suhu (Z_2) 318.41326531, 350.20918367 dan 358.15816327 dan variabel trombosit (Z_3) 14.55170068, 14.98707483 dan 15.09591837.

Pemilihan Titik Knot Terbaik

Titik knot terbaik merupakan titik knot yang mempunyai nilai GCV dan MSE minimum. Berikut merupakan perbandingan nilai GCV dan MSE minimum yang diperoleh pada satu titik knot, dua titik knot, dan tiga titik knot yang ditunjukkan pada Tabel 3.

TABEL 3. Perbandingan Nilai GCV dan MSE

Model	GCV	MSE
1 Titik Knot	244.82556133	204.4337
2 Titik Knot	244.82556133	204.4337
3 Titik Knot	221.67745153	199.1032

Berdasarkan kriteria pemilihan model terbaik diketahui bahwa nilai GCV dan MSE paling minimum dihasilkan oleh model regresi nonparametrik spline dengan tiga titik knot.

Pengujian Parameter Model

Pengujian parameter model dilakukan dengan serentak (simultan) kemudian dilanjutkan dengan pengujian secara parsial (individu).

TABEL 4. Uji serentak estimasi model regresi semiparametrik spline

Sumber Keragaman	Derajat Bebas	Jumlah Kuadrat (JK)	Rataan Jumlah Kuadrat (RJK)	F_{hitung}	F_{tabel} $\alpha = 5\%$
Regresi	12	11885,33	990,4442	22,4788	1,97
Residual	45	1982,754	44,0612		
Total	57	13868,09			

Dengan taraf nyata $\alpha = 5\%$ diperoleh $F_{hitung} \geq F_{tabel}$ yaitu 22,4788≥ 1,97 maka H_0 ditolak. Hal ini berarti terdapat pengaruh yang signifikan secara bersama-sama antara variabel bebas terhadap variabel terikat, sehingga model signifikan.

Pada Tabel 5 menjelaskan bahwa dua variabel prediktor mempunyai parameter yang signifikan terhadap model karena memiliki *p-value* ≤ 5%. Maka variabel hemoglobin (X) dan trombosit (Z_3) berpengaruh secara signifikan terhadap lama kesembuhan pasien.

TABEL 5. Uji individu estimasi model regresi semiparametrik spline

Variabel	Parameter	Estimasi	$p - value$	Keterangan
	β_0	332,588	0,003	Signifikan
X	β_1	0,227	0,000	Signifikan
Z_1	β_2	0,012	0,058	Tidak signifikan
	β_3	-7,827	0,617	Tidak signifikan
	β_4	0,379	0,840	Tidak signifikan
	β_5	0,008	0,993	Tidak signifikan
Z_2	β_6	-0,056	0,503	Tidak signifikan
	β_7	0,004	0,438	Tidak signifikan
	β_8	0,119	0,010	Signifikan
	β_9	0,002	0,819	Tidak signifikan
Z_3	β_{10}	-0,004	0,000	Signifikan
	β_{11}	2,974	0,682	Tidak signifikan
	β_{12}	-8,745	0,012	Signifikan
	β_{13}	-0,41	0,757	Tidak signifikan

Pengujian Residual Model

Residual (*goodness of fit*) dari suatu model regresi harus memenuhi asumsi $IIDN(0, \sigma^2)$ yaitu identik, independen, dan berdistribusi normal. Uji homogenitas dilakukan dengan menggunakan uji Glejser. Pada uji homogenitas diperoleh nilai F_{hit} sebesar 0,000 dengan signifikansi $\alpha = 5\%$, F_{tabel} sebesar 2,55. Nilai $F_{hit} \leq F_{tabel}$ yang mengindikasikan H_0 diterima. Maka dapat disimpulkan bahwa semua variabel tidak berpengaruh signifikan terhadap nilai mutlak residual.

Uji asumsi independen dilakukan dengan menggunakan uji d Durbin-Watson. Pada uji asumsi independen diperoleh nilai d sebesar 2,027. Selanjutnya nilai d dibandingkan dengan nilai tabel Durbin-Watson dengan signifikasi $\alpha = 5\%$, jumlah sampel sebanyak 58 (T=58), k = 5 yakni satu variabel dependen dan empat variabel independen. Dari tabel Durbin-Watson diperoleh nilai d_L (batas bawah Durbin-Watson) dan d_U (batas atas Durbin-Watson) secara berturut-turut yaitu 1,3953 dan 1,7673.

$$4\text{-}d_L = 4\text{-}1{,}3953 = 2{,}6047$$

$4\text{-}d_U = 4\text{-}1{,}7673 = 2{,}2327$

Karena nilai d Durbin-Watson terletak diantara nilai d_U dan $4 - d_U$, berdasarkan Tabel 1 aturan keputusan Durbin-Watson, maka H_0 diterima. Hal ini mengindikasikan bahwa tidak terdapat autokorelasi positif ataupun negatif pada residual.

Uji normalitas dilakukan dengan melakukan uji Anderson-Darling. Pada uji normalitas diperoleh nilai Anderson-Darling sebesar 0,384 dan $p - value$ sebesar 0,384 dengan signifikansi 5%. Karena $p - value \geq 0{,}05$, maka H_0 diterima. Hal ini mengindikasikan bahwa residual model memenuhi asumsi distribusi normal.

Dengan demikian dapat disimpulkan bahwa residual dari model regresi semiparametrik spline linear dengan tiga titik knot memenuhi asumsi $IIDN(0, \sigma^2)$ yaitu identik, independen, dan berdistribusi normal.

Koefisien Determinasi

Nilai koefisien determinasi (R^2) menunjukkan seberapa besar kebaikan model regresi dalam menjelaskan keragaman lama kesembuhan pasien demam berdarah dengue. Diperoleh nilai R^2 sebesar 85,7% berarti model regresi semiparametrik spline yang didapatkan mampu menjelaskan keragaman lama kesembuhan pasien demam berdarah dengue. Nilai tersebut mendekati 100%, sehingga model sudah cukup baik.

Interpretasi Model Regresi Semiparametrik *Spline* dengan Tiga Titik Knot

Estimasi model yang diperoleh adalah sebagai berikut :

$$\hat{y} = 332{,}588 + 0{,}227x - 0{,}004z_3 + 2{,}974(z_3 - 14{,}552) - 8{,}745(z_3 - 14{,}987) - 0{,}41(z_3 - 15{,}096)$$

Berdasarkan model tersebut, maka dapat diinterpretasikan sebagai berikut :

Apabila variabel hemoglobin (X) dinggap konstan, maka pengaruh trombosit (Z_3) terhadap lama kesembuhan pasien adalah

$$\hat{y} = 332{,}588 - 0{,}004z_3 + 2{,}974(z_3 - 14{,}552) - 8{,}745(z_3 - 14{,}987) - 0{,}41(z_3 - 15{,}096)$$

$$= \begin{cases} 332{,}588 - 0{,}004z_3 & ; z_3 < 14{,}552 \\ 289{,}310 + 2{,}97z_3 & ; 14{,}552 \leq z_3 < 14{,}987 \\ 463{,}649 - 8{,}749z_3 & ; 14{,}987 \leq z_3 < 15{,}096 \\ 338{,}777 - 0{,}414z_3 & ; z_3 \geq 15{,}096 \end{cases}$$

Ketika jumlah trombosit kurang dari 14,552 maka y akan mengalami penurunan sebesar 0,004. Apabila jumlah trombosit berada diantara 14,552 dan 14,987 maka y akan mengalami kenaikan 2,97. Apabila jumlah trombosit berada diantara 14,987 dan 15,096, maka y akan mengalami penurunan sebesar 8,749. Dan apabila kadar trombosit lebih dari 15,096, maka y akan mengalami penurunan sebesar 0,003. Koefisien bernilai negatif artinya terjadi hubungan negatif antara trombosit dengan y yang mengindikasikan bahwa apabila kadar trombosit meningkat menyebabkan kesembuhan pasien cenderung semakin cepat dan begitupula sebaliknya.

Apabila trombosit (Z_3) dianggap konstan, maka interpretasi terhadap variabel hemoglobin adalah apabila hemoglobin mengalami kenaikan 1(g/dL) maka y akan mengalami kenaikan sebesar 0,227. Koefisien bernilai positif artinya terjadi hubungan positif antara hemoglobin dengan y. Hal ini mengindikasikan bahwa apabila jumlah hemoglobin meningkat maka berakibat pada jenjang waktu yang lebih lama pada kesembuhan pasien.

PEMBAHASAN

Penelitian tentang model regresi nonparametrik untuk data longitudinal telah dilakukan oleh Utami (2014). Penelitian ini mengkaji estimator polinomial lokal kernel dalam pemodelan regresi nonparametrik untuk data longitudinal. Penelitian Abdy (2009) dan Yani (2017) memodelkan regresi semiparametrik. Abdy (2009) menerapakan pendugaan parameter *Generalized Estimating Equation* dalam memodelkan regresi semiparametrik. Yani (2017) menerapkan *spline truncated linear* dalam pemodelan regresi semiparametrik.

Sedangkan pada penelitian ini mengkaji model regresi semiparametrik *spline* dengan pendekatan *Generalized Estimating Equation* untuk data longitudinal. Hasil dari penelitian ini diperoleh model regresi semiparametrik yaitu sebagai berikut :

$$\hat{y} = 332{,}588 + 0{,}227x - 0{,}004z_3 + 2{,}974(z_3 - 14{,}552) - 8{,}745(z_3 - 14{,}987) - \\ -0{,}41(z_3 - 15{,}096)$$

dengan nilai GCV minimum 221.67745153**,** nilai MSE sebesar 199.1032**,** nilai koefisien determinasi sebesar 85,7%.

KESIMPULAN

Estimasi model regresi semiparametrik dengan kriteria nilai GCV minimum pada model regresi semiparametrik *spline* dengan tiga titik knot adalah :

$$\hat{y} = 332{,}588 + 0{,}227x - 0{,}004z_3 + 2{,}974(z_3 - 14{,}552) - 8{,}745(z_3 - 14{,}987) - \\ 0{,}41(z_3 - 15{,}096)$$

dengan nilai GCV minimum 221.67745153**,** nilai MSE sebesar 199.1032 yang dicapai pada titik knot 14,552; 14,987; 15,096. Koefisien determinasi sebesar 85,7% keragaman lama kesembuhan pasien Demam Berdarah Dengue (DBD) yang menjalani rawat inap di Rumah Sakit Unhas Makassar.

Pada penelitian ini dibahas model regresi semiparametrik dengan pendugaan parameter GEE untuk satu variabel pada komponen parametriknya. Sehingga penelitian lebih lanjut dapat dilakukan untuk lebih dari satu variabel komponen parametrik dan menggunakan pendugaan parameter yang berbeda.

DAFTAR PUSTAKA

Abdy, M. (2009). Regresi Semiparametrik dengan Pendekatan Generalized Estimating Equation (GEE). *Jurnal Matematika, Statistika dan Komputasi*, *5*(2). 66-75.

Adawiyah, R. (2018). *Model Regresi Nonparametrik dengan Pendekatan Spline (Studi Kasus : Berat Badan Lahir Rendah di Rumah Sakit Ibu dan Anak Siti Fatimah Makassar*). Universitas Negeri Makassar, Makassar.

Danardono. (2015). *Analisis Data Longitudinal*. Yogyakarta : UGM Press.

Gusti, O.W. (2011). *Regresi Semiparametrik Spline Dalam Memodelkan Hasil UNAS SMAN 1 Sekaran Lamongan*. Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang.

Handayani, L & Putera, F.H.A. (2016). Model Data Longitudinal dengan Pendekatan Generalized Estimating Equation pada Struktur Korelasi Exchangeable, Auto-Regressive, dan Unstructured. *Jurnal Statistika Universitas Tadulako,16*(1). 4-16.

Ibrahim, N.A. & Suliadi. (2009). Nonparametric Regression for Correlated Data. *Article in WSEAS Transaction on Mathemathics, 8*(7). 1109-2769.

Laome, L. (2009). Model Regresi Semiparametrik Spline Untuk Data Longitudinal pada Kasus Kadar CD4 Penderita HIV. *Jurnal Matematika Murni dan Terapan, 13*(2). 189-194.

Poerwanto, B & Budiantara, I. N. (2014). Estimasi Kurva Regresi Semiparametrik Spline Untuk Data Longitudinal. *Prosiding Seminar Nasional Matematika.* Bali, Indonesia: Universitas Udayana.

Utami, T. W. (2014). Pemodelan Regresi Nonparametrik pada Data Longitudinal Berdasarkan Estimasi Polonomial Lokal Kernel. *Jurnal Statistika, 5*(2).602-610.

Yani, N. W. M. N. (2017). Aplikasi Model Regresi Semiparametrik Spline Truncated *Jurnal Matematika, 6*(1). 65-73.